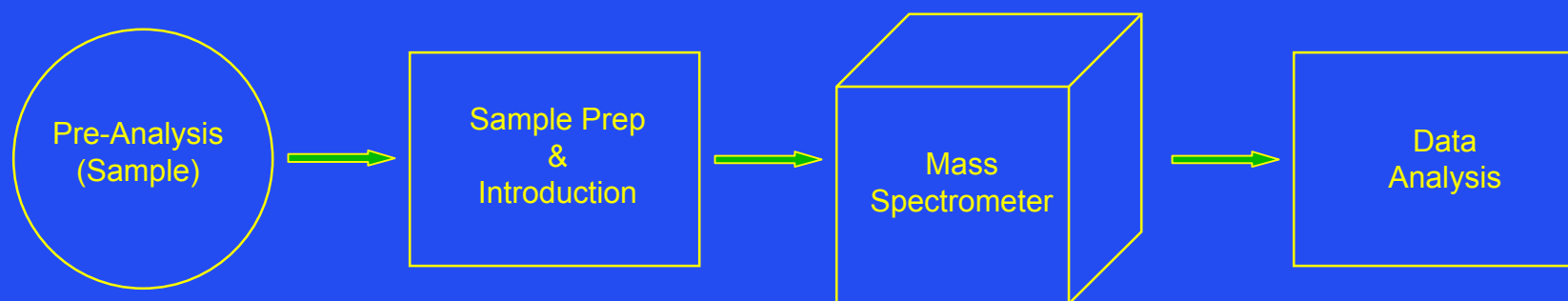


Spectral Variability - What We Might Do With Better Mass Spectra

Alfred L Yergey
NICHD, NIH

Sources of Analytical Variability for Mass Spectrometric Studies



Population Issues

- Diet, Age, Gender, Medication, etc.

Sample Acquisition

- Tubes, Storage, etc.

Experimental Design

- Replicates
- Run Order
- Normal Error?

One Consequence of Poor Experimental Design

- Bad mass spectrometry
- Unsound / poorly executed mathematical analysis
- Over-interpretation of results

MECHANISMS OF DISEASE

Mechanisms of disease

Use of proteomic patterns in serum to identify ovarian cancer

Emanuel F Petricoin III, Ali M Ardekani, Ben A Hitt, Peter J Levine, Vincent A Fusaro, Seth M Steinberg, Gordon B Mills, Charles Simone, David A Fishman, Elise C Kohn, Lance A Liotta

Summary

Background New technologies for the detection of early-stage ovarian cancer are urgently needed. Pathological changes within an organ might be reflected in proteomic patterns in serum. We developed a bioinformatics tool and used it to identify proteomic patterns in serum that distinguish neoplastic from non-neoplastic disease within the ovary.

Methods Proteomic spectra were generated by mass spectroscopy (surface-enhanced laser desorption and ionisation). A preliminary "training" set of spectra derived from analysis of serum from 50 unaffected women and 50 patients with ovarian cancer were analysed by an iterative searching algorithm that identified a proteomic pattern that completely discriminated cancer from non-cancer. The discovered pattern was then used to classify an independent set of 116 masked serum samples: 50 from women with ovarian cancer, and 66 from unaffected women or those with non-malignant disorders.

Findings The algorithm identified a cluster pattern that, in the training set, completely segregated cancer from non-cancer. The discriminatory pattern correctly identified all 50 ovarian cancer cases in the masked set, including all 18 stage I cases. Of the 66 cases of non-malignant disease, 63 were recognised as not cancer. This result yielded a sensitivity of 100% (95% CI 93–100), specificity of 95% (87–99), and positive predictive value of 94% (84–99).

Interpretation These findings justify a prospective population-based assessment of proteomic pattern technology as a screening tool for all stages of ovarian cancer in high-risk and general populations.

Lancet 2002; 359: 572–77

Food and Drug Administration/National Institutes of Health Clinical Proteomics Program, Department of Therapeutic Proteins/Center for Biologics Evaluation and Research, Food and Drug Administration, Bethesda, MD, USA (E F Petricoin *m*), A M Ardekani *m*); Laboratory of Pathology (L A Liotta *sc*), E C Kohn *sc*, V A Fusaro) and Biostatistics and Data Management Section, Center for Cancer Research (S M Steinberg *m*), National Cancer Institute, National Institutes of Health, Bethesda, MD; Correlogic Systems Inc, Bethesda, MD (B A Hitt *m*, P J Levine *sc*); Department of Molecular Therapeutics, Division of Cancer Medicine, M D Anderson Cancer Center, Houston, TX (G B Mills *sc*); Simone Protective Cancer Institute, Lawrenceville, NJ (C B Simone *sc*); and National Ovarian Cancer Early Detection Program, Northwestern University Medical School, Chicago, IL (D A Fishman *sc*).

Correspondence to: Dr Emanuel F Petricoin III, Building 29A, Room 2B02, 8800 Rockville Pike, Bethesda, MD 20892, USA (e-mail: petricoin@cber.fda.gov).

Introduction

Application of new technologies for detection of ovarian cancer could have an important effect on public health,¹ but to achieve this goal, specific and sensitive molecular markers are essential.^{2–4} This need is especially urgent in women who have a high risk of ovarian cancer due to family or personal history of cancer, and for women with a genetic predisposition to cancer due to abnormalities in predisposition genes such as *BRCA1* and *BRCA2*. There are no effective screening options for this population.

Ovarian cancer presents at a late clinical stage in more than 80% of patients,⁵ and is associated with a 5-year survival of 35% in this population. By contrast, the 5-year survival for patients with stage I ovarian cancer exceeds 90%, and most patients are cured of their disease by surgery alone.^{1–4} Therefore, increasing the number of women diagnosed with stage I disease should have a direct effect on the mortality and economics of this cancer without the need to change surgical or chemotherapeutic approaches.

Cancer antigen 125 (CA125) is the most widely used biomarker for ovarian cancer.^{1–4} Although concentrations of CA125 are abnormal in about 80% of patients with advanced-stage disease, they are increased in only 50–60% of patients with stage I ovarian cancer.^{1–4} CA125 has a positive predictive value of less than 10% as a single marker, but the addition of ultrasound screening to CA125 measurement has improved the positive predictive value to about 20%.⁶

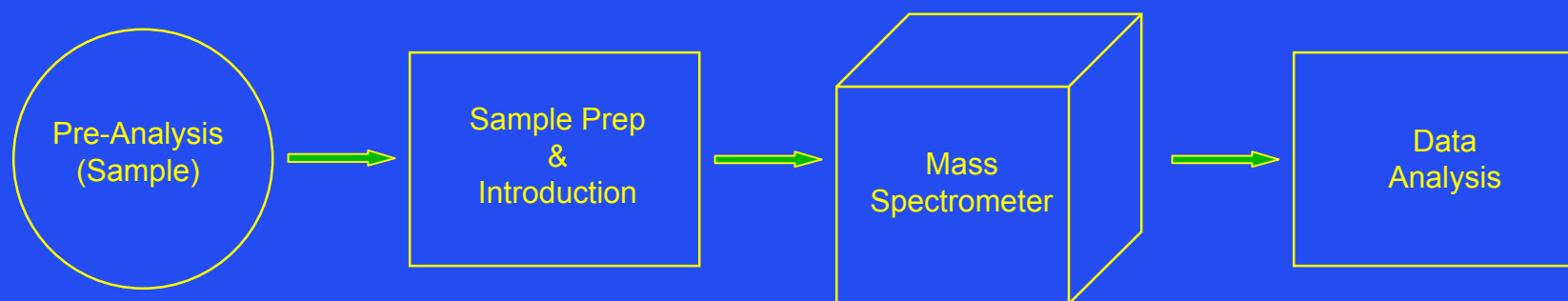
Low-molecular-weight serum protein profiling might reflect the pathological state of organs and aid in the early detection of cancer. Matrix-assisted laser desorption and ionisation time-of-flight (MALDI-TOF) and surface-enhanced laser desorption and ionisation time-of-flight (SELDI-TOF) mass spectroscopy can profile proteins in this range.^{7–9} These profiles can contain thousands of data points, necessitating sophisticated analytical tools. Bioinformatics has been used to study physiological outcomes and cluster gene microarrays,^{10–12} but to uncover changes in complex mass spectrum patterns of serum proteins, higher order analysis is required. We aimed to link SELDI-TOF spectral analysis with a high-order analytical approach using samples from women with a known diagnosis to define an optimum discriminatory proteomic pattern. We then aimed to use this pattern to predict the identity of masked samples from unaffected women, women with early-stage and late-stage ovarian cancer, and women with benign disorders.

Participants and methods

Study population

100 control samples (50 for the preliminary analysis and 50 for the masked analysis) were provided from the National Ovarian Cancer Early Detection Program (NOCEDEP) clinic at Northwestern University

Sources of Analytical Variability for Mass Spectrometric Studies



Analytical Approach

- Single band or shotgun

Digestion conditions

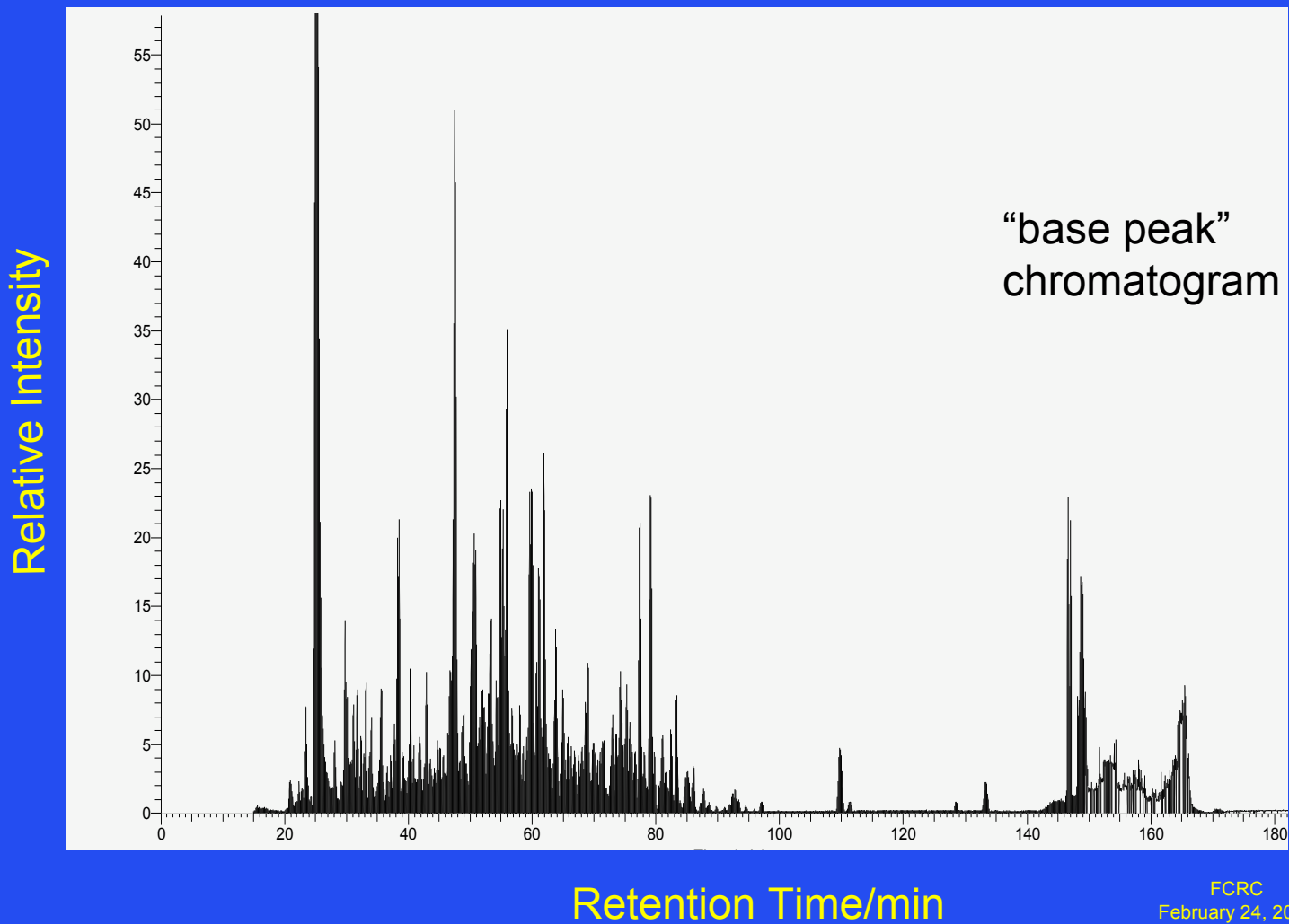
- Enzyme
- Time

Columns

- Solid phase
- Gradient & solvent

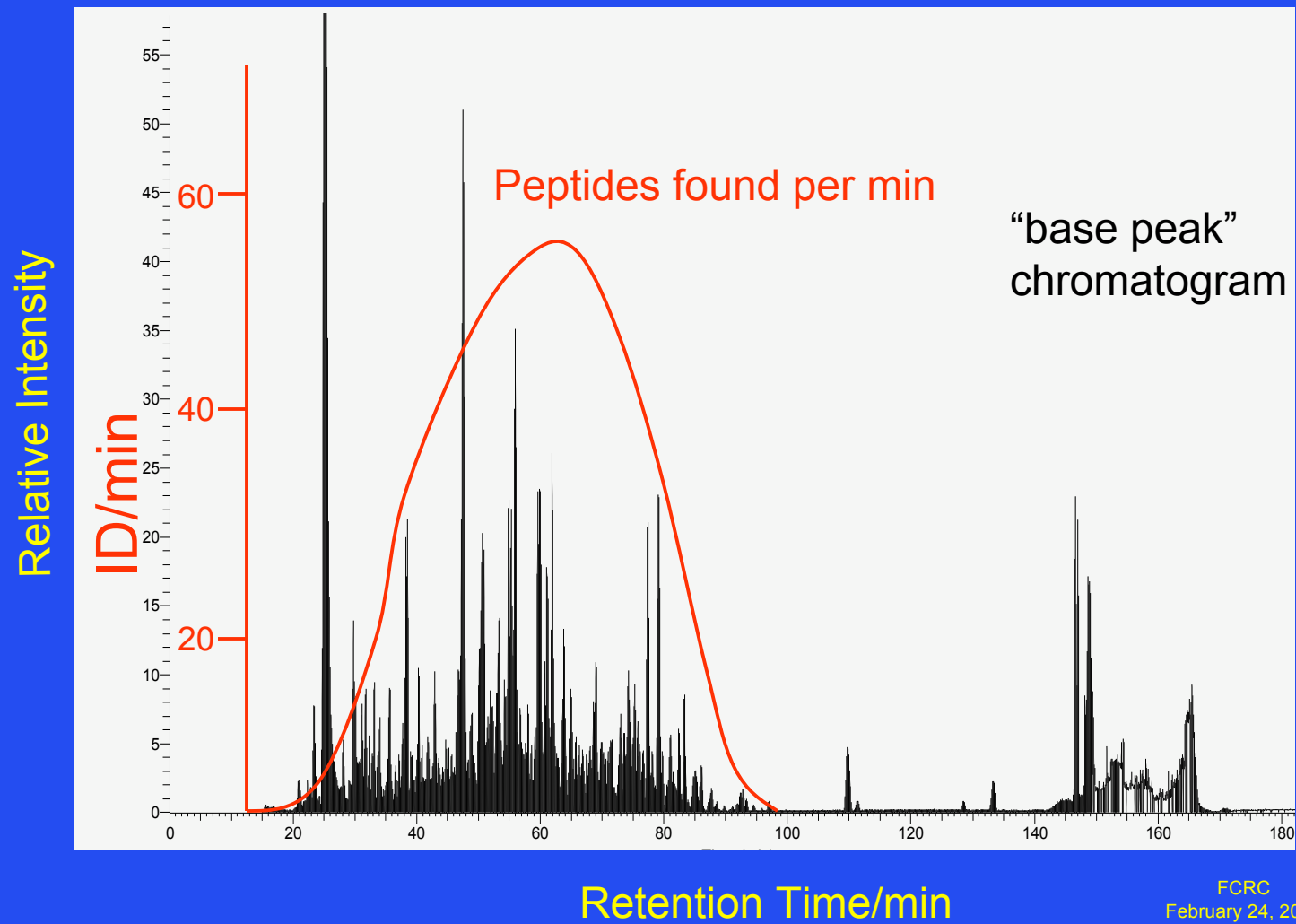
Source conditions

Chromatograph (yeast lysate)



FCRC
February 24, 2009

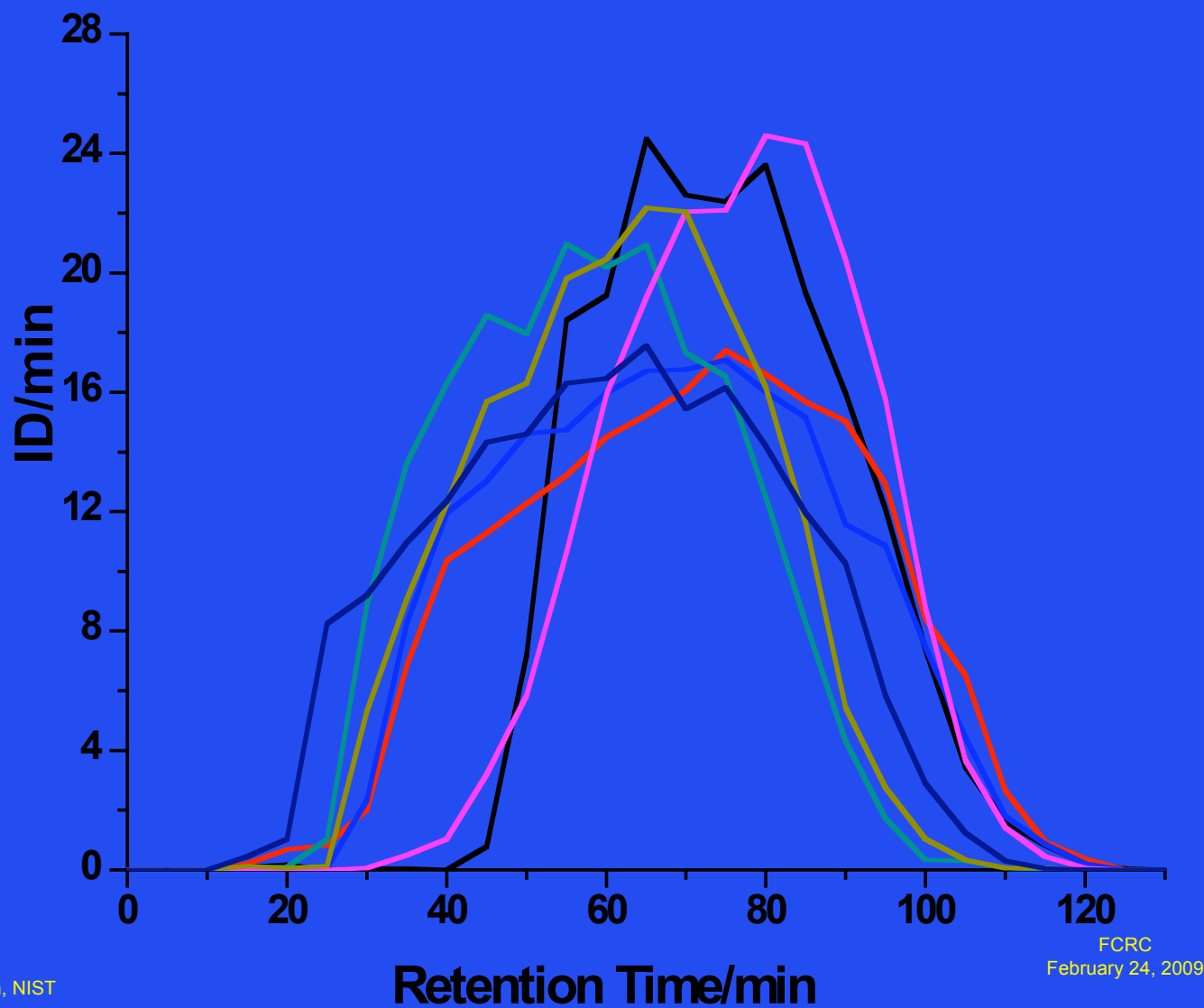
Chromatograph (yeast lysate)



Shotgun Proteomics

- Digest proteins to peptides
 - Separate peptides (Cation * C-18 LC)
 - Identify peptides (from CID fragments)
 - Infer proteins
-
- Hundreds of IDed peptides
 - bad news: differ run-to-run
 - good news: great for QA/QC!

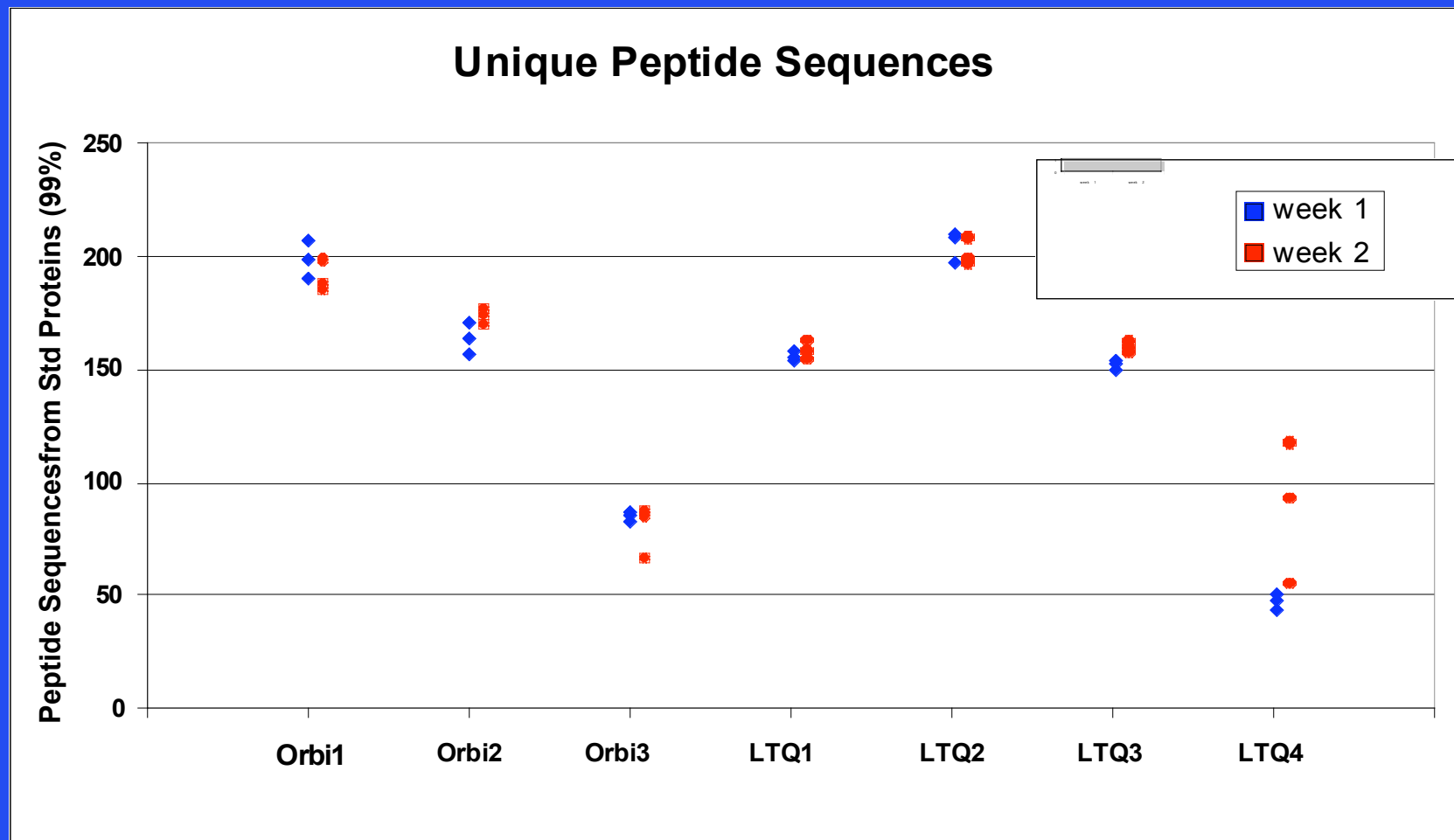
7 Labs, yeast, simple SOP



FCRC
February 24, 2009

CPTAC Study 2*

3 Major Plasma Proteins



* Liebler et al, ThP 602

FCRC
February 24, 2009

Characterizing the Variability

A new suite of programs can be used to look at the variability in LC MS/MS data:

An attempt to standardize ESI conditions is being made with the use of 'Thermometer Ions' by John Peltier and colleagues:

NISTMSQC1

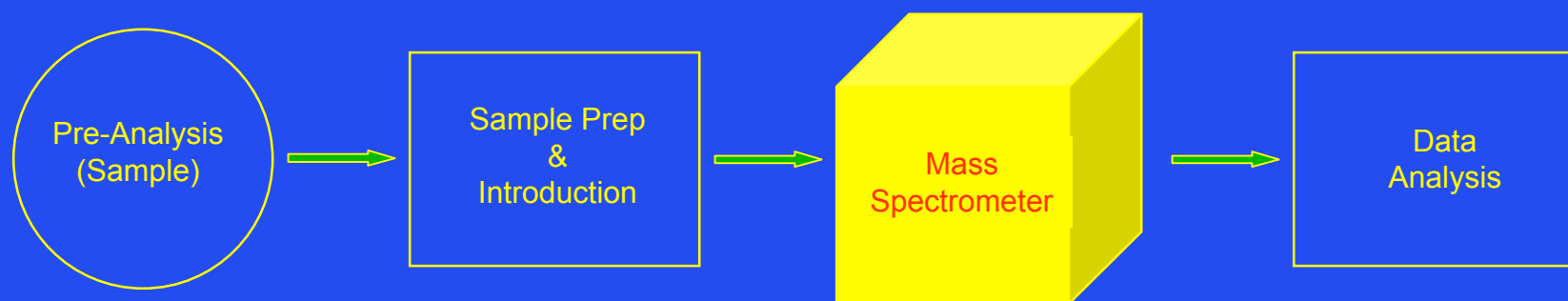
Contact

Paul Rudnick or Steve Stein

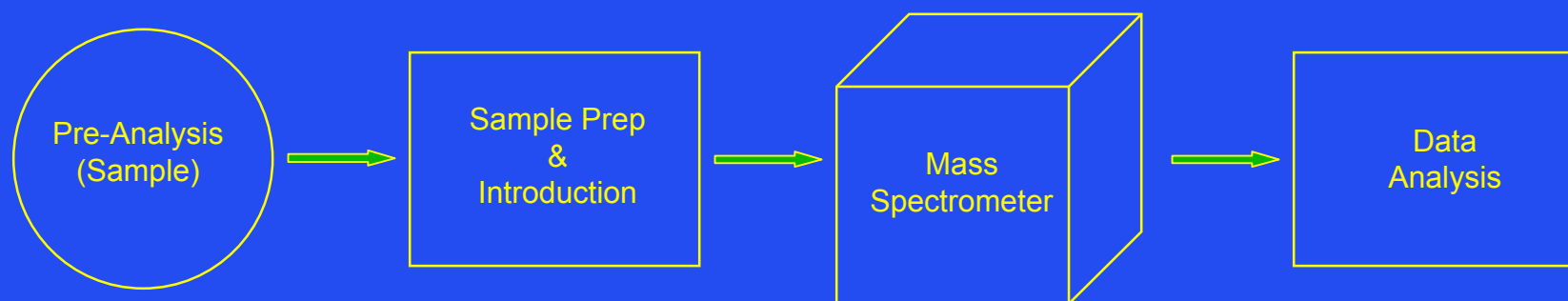
At NIST

Defining Instrument Performance and Assessing the Reproducibility of Mass Spectrometric Analyses of Complex Samples - TPM 371

Sources of Analytical Variability for Mass Spectrometric Studies



Sources of Analytical Variability for Mass Spectrometric Studies



Search Engine

- “Significance”
- FP rate

Data Base version

Coverage

PTMs

REPLICATES

Approach for Analysis of ABRF Sample

Sample Preparation for MS/MS Analysis

Dissolve sample in 0.1% SDS/0.1 M AmHCO₃
Reduce & alkylate Cys with iodoacetamide
Bring up in SDS/PAGE sample buffer,
separate on gel, and stain.

Cut out bands & digest with trypsin O/N.

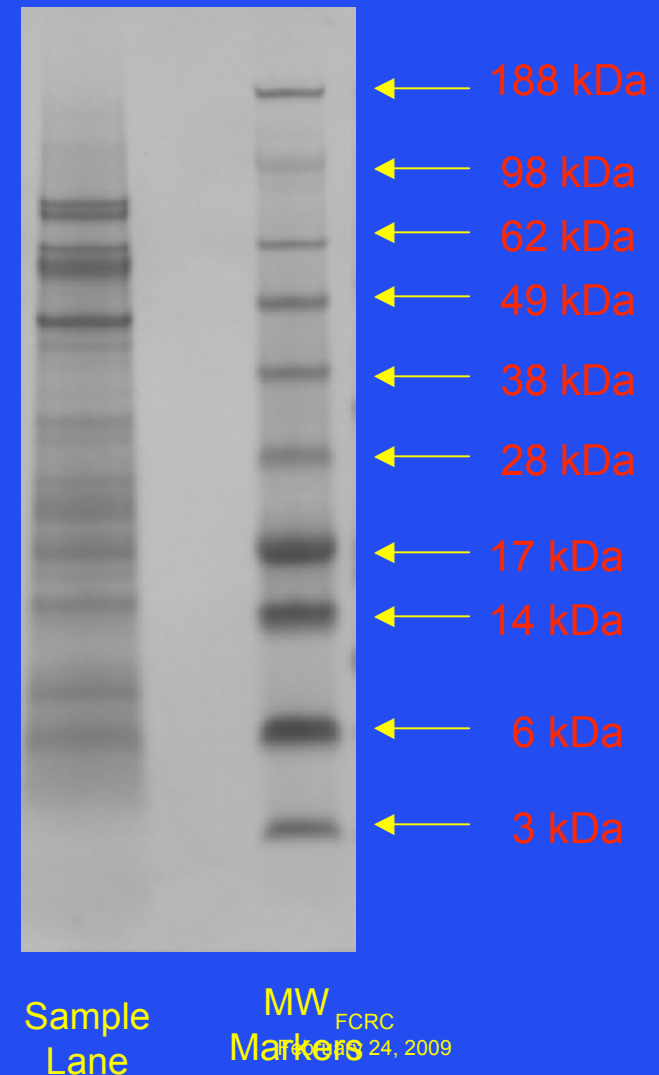
Extract peptides.

MS/MS Analysis

LC/MS/MS
(ion-trap)

Off-line cap. RP HPLC

MALDI MS/MS
(TOF-TOF)



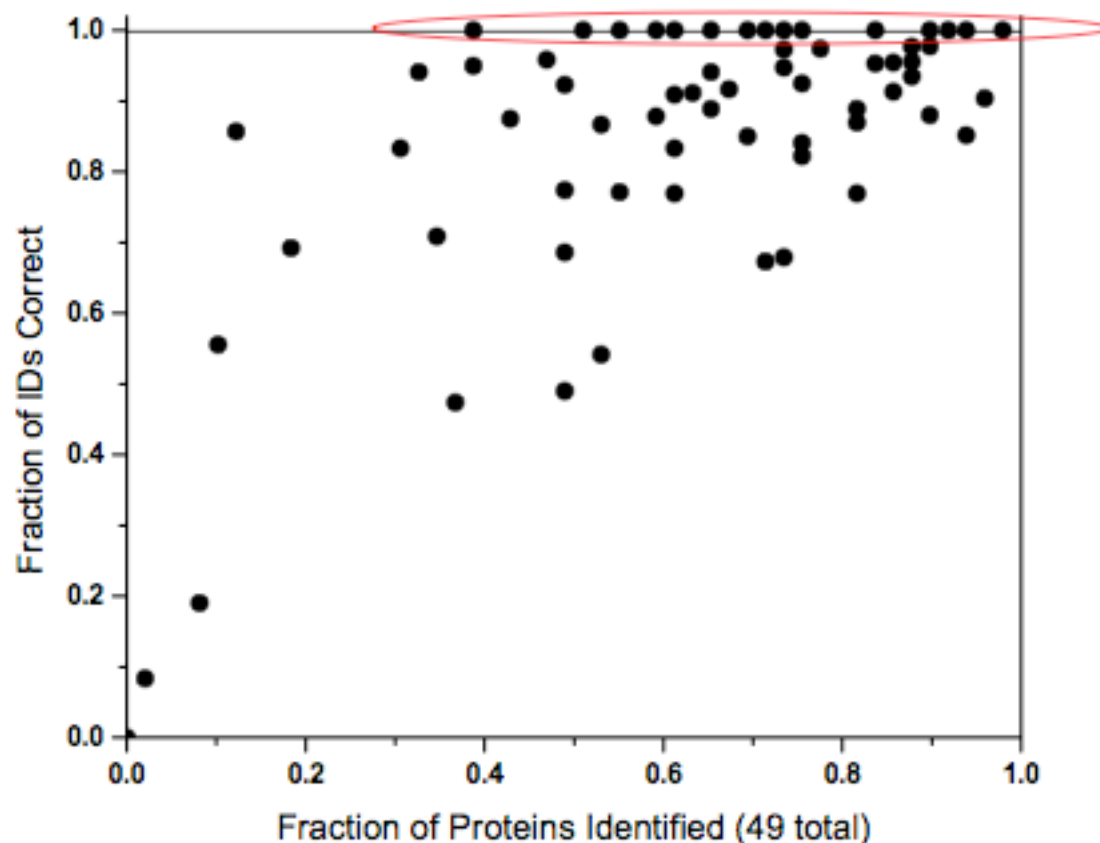
Summary of Protein Identification Results SMSM, NICHD

- Number of Proteins “Identified”
 - 36 from sample (73% of 49 proteins in sample)
 - 4 contaminants
 - 0 Incorrect (false positive)
- > 4 peptides/protein found for 34/36 proteins

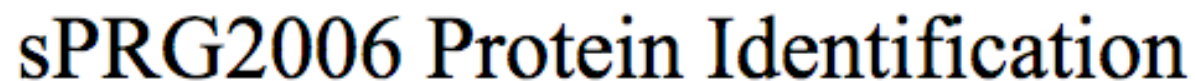


*Proteomics Standards
Research Group*

Protein Identification Performance for 74 Labs



The performance of each lab is represented by a point defined by the fraction of all known proteins identified and the fraction of all reported identifications that were correct. Those labs with no false identifications fall along the upper axis.

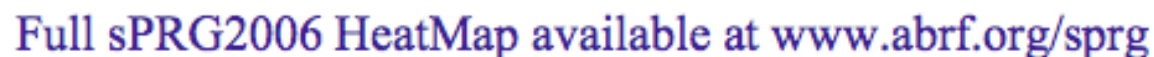


Legend:

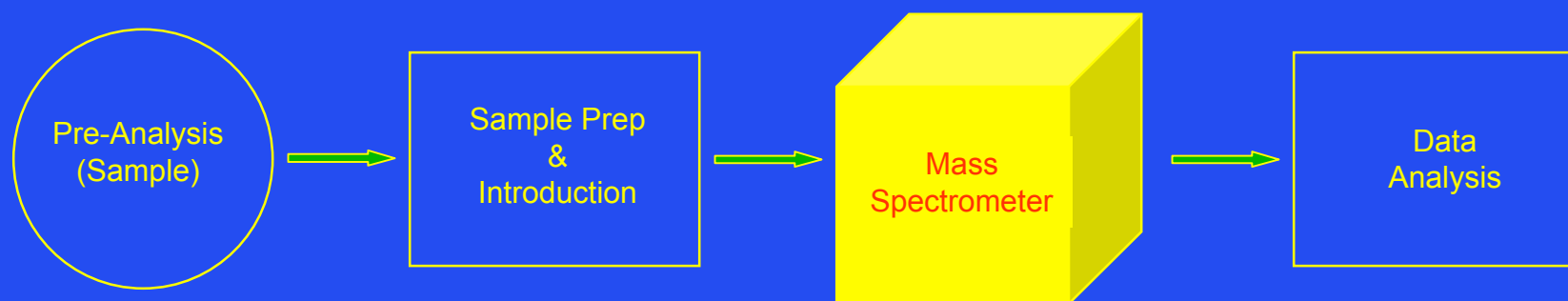
1 - 2

3 - 5

> 5



Sources of Analytical Variability for Mass Spectrometric Studies

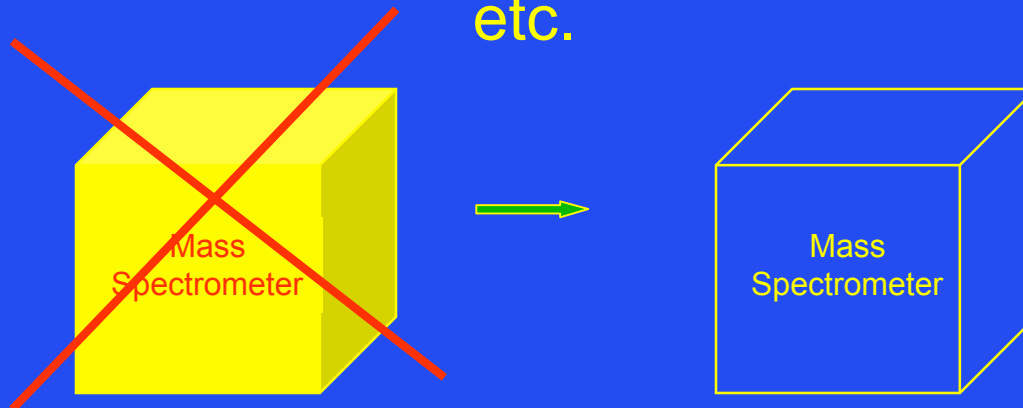


Where does this take us?

Perhaps -

We need to recognize that mass spectrometers need
to be used to collect mass spectra -

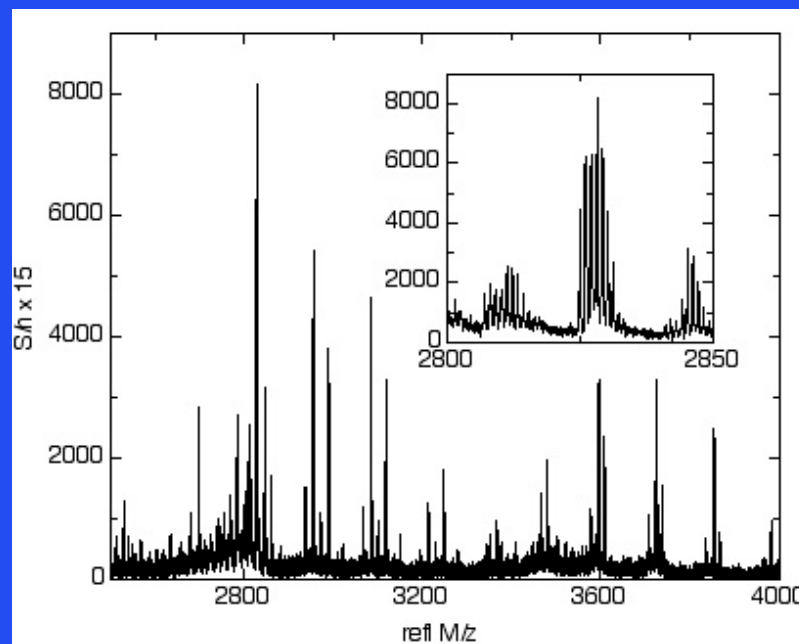
and NOT simply to generate
“Identifications”
“Biomarkers”
etc.



Why Worry About These Issues?

In generating complex spectra from different sources -
how do you tell if they are similar or different?

- How does one go about comparing different complex spectra? (Other than by holding them up to a window?)
- How does one identify the most reproducible features of spectra when multiple (discordant) replicates are available?
- **We need an automated and robust method to compare replicates and differentiate spectra from a various sources.**



General Method for Applications

- Produce multiple replicates of a MALDI spectrum
- Generate a Consensus Spectrum from the overall mean
- Use the *Dot Product and its Confidence Interval* to eliminate poor replicates
- Compare Consensus Spectra from different samples and assess similarity using the DP and CI.

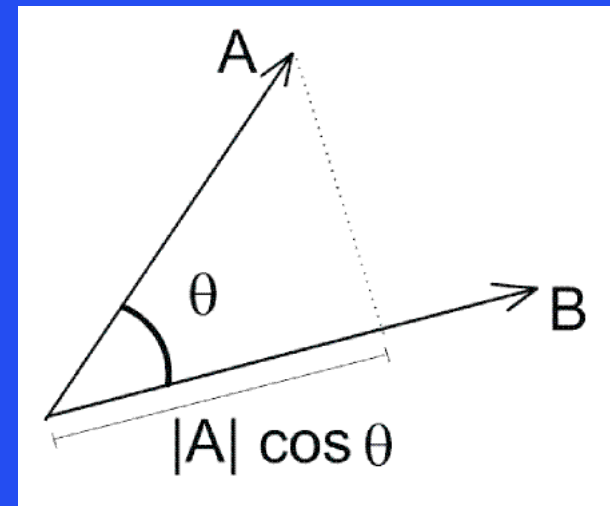
What IS the Dot Product?

- In Euclidian space, the dot product of 2 vectors is given by:

$$A \cdot B = |A||B|\cos \theta$$

where θ is the angle between the vectors.

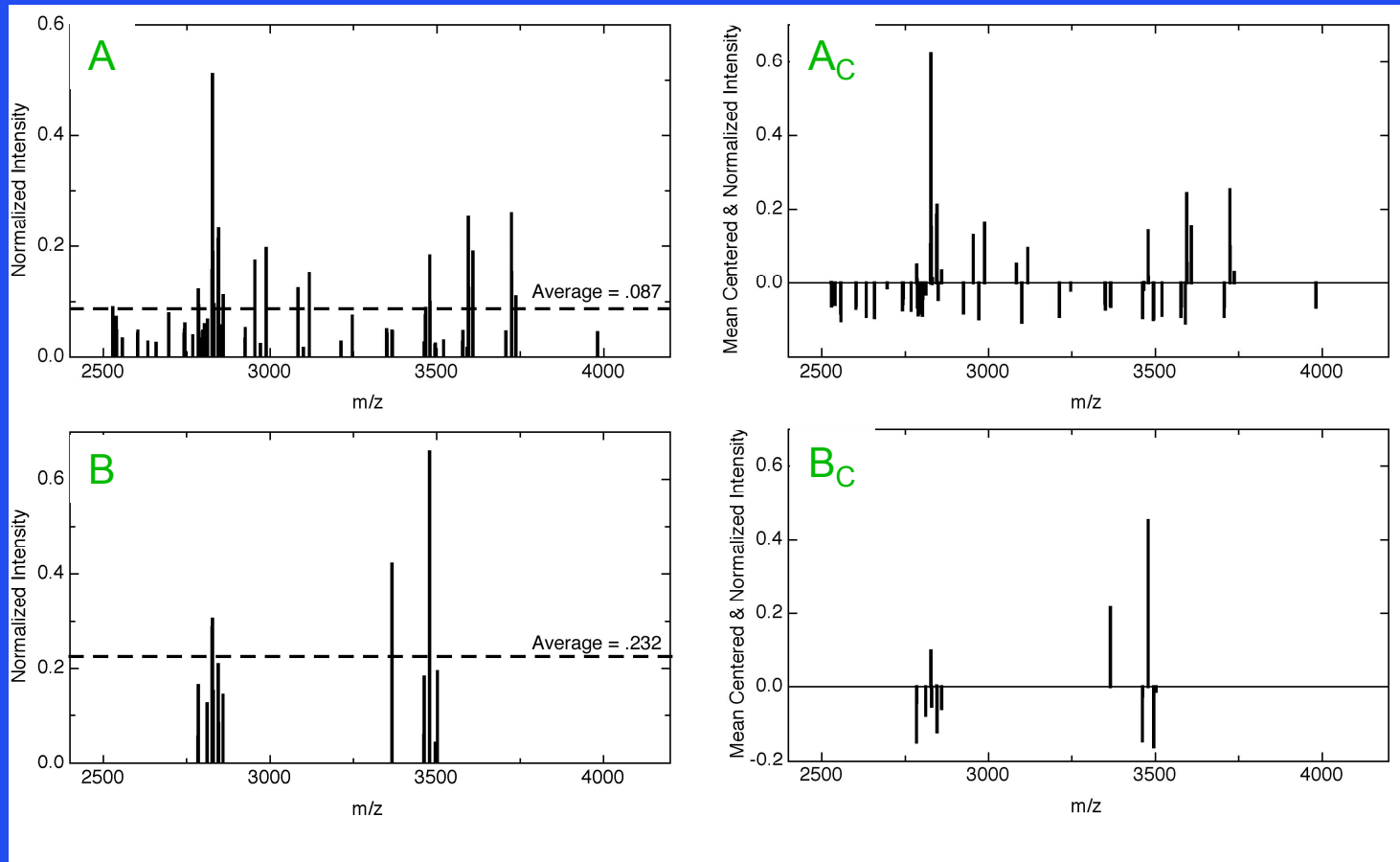
- Since the cosine of $0 = 1$, the closer 2 vectors are to being parallel or overlapping, the closer their dot product is to 1.



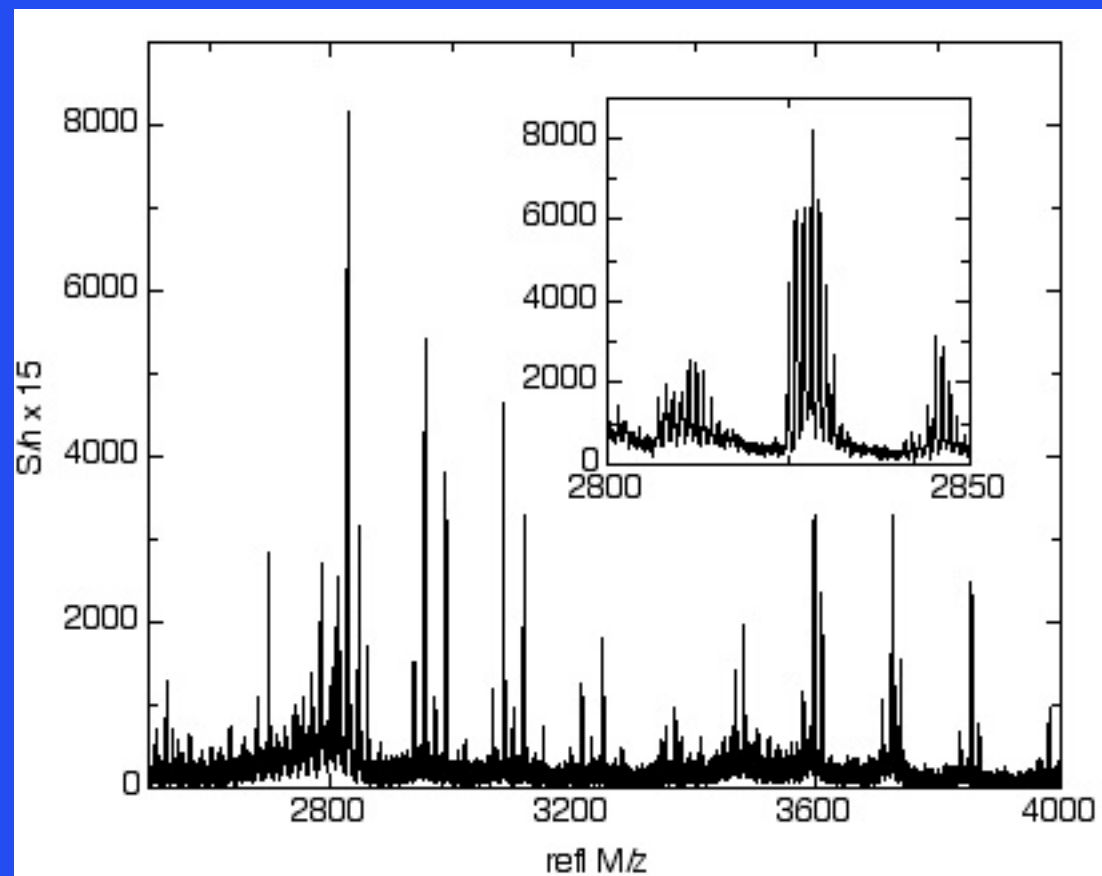
When the normalized vectors are
mean centered

the dot product *IS IDENTICAL TO*
Pearson's Correlation Coefficient

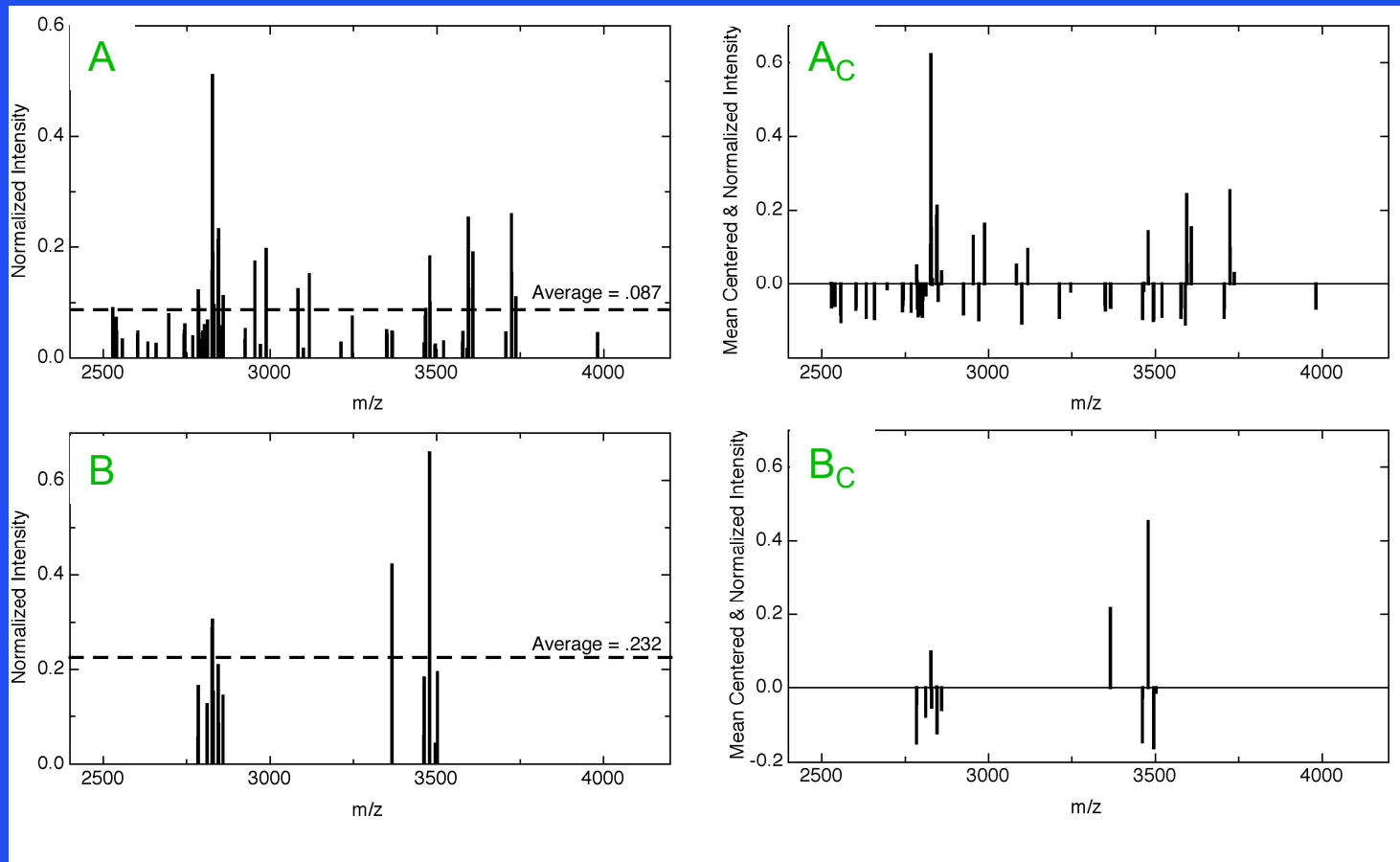
Mean Centering



Typical Reflector Spectrum of Rat Brain Tubulin



Mean Centering



$DP(AB) = 0.46$

$DP(A_C B_C) = 0.19$

95% CI = (-0.46 to 0.74)

February 24, 2009

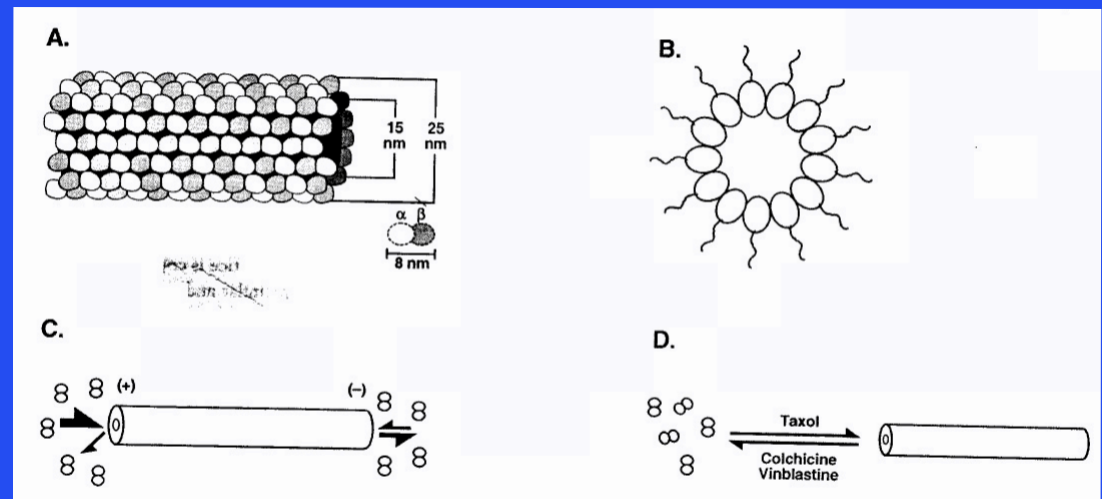
Tubulin Post-translational Modifications

Why Study Tubulins?

What Are They?

Tubulins are ~50 kDa proteins that polymerize into microtubules and are involved with:

- Intracellular transport
- Ciliary function
- Mitosis



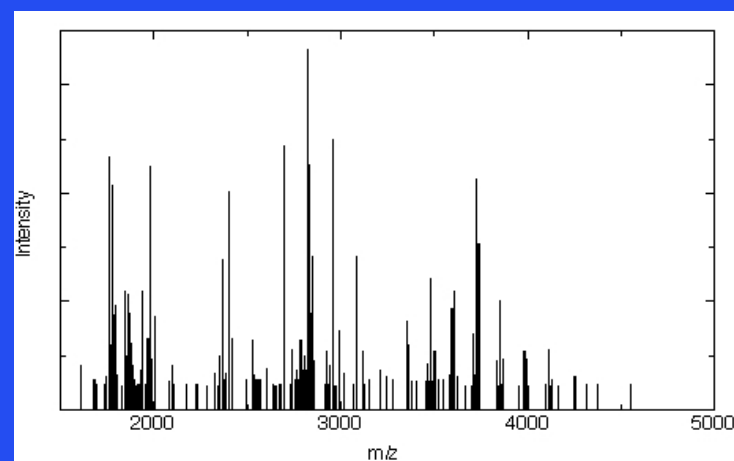
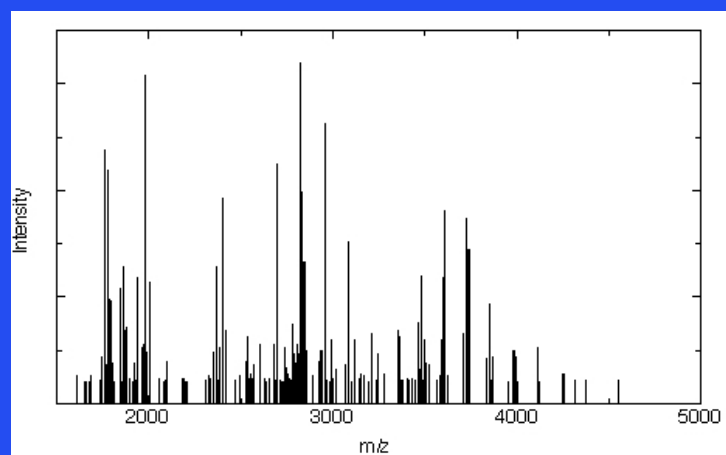
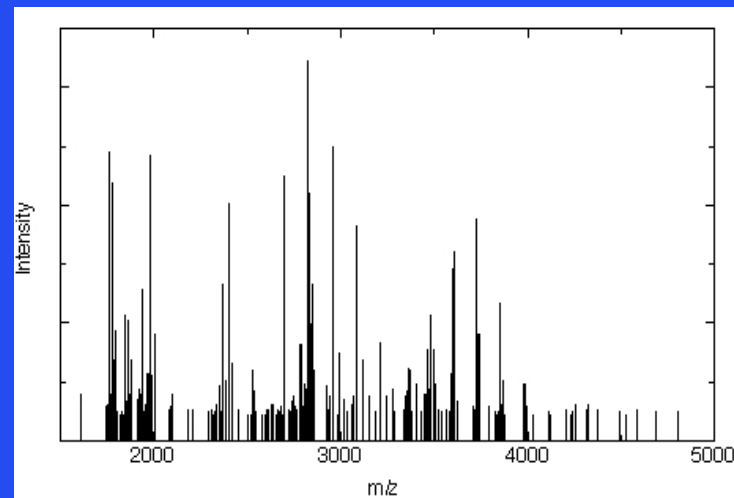
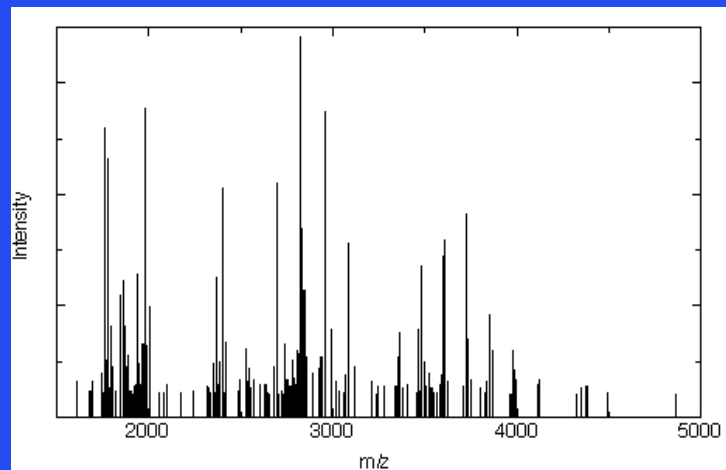
Determination of C-termini

- CNBr cleavage (cuts on C-term of Met)
- Negative ion reflector spectra were obtained using an ABI 4800 TOF-TOF
- Each sample was spotted in triplicate and 10 replicate spectra, 1000 shots each, were obtained from each spot.
- All spectra were calibrated externally using ChET

Nature of the Problem - Example 1

Replicate Spectra from Rat Brain Tubulin

All intensities normalized



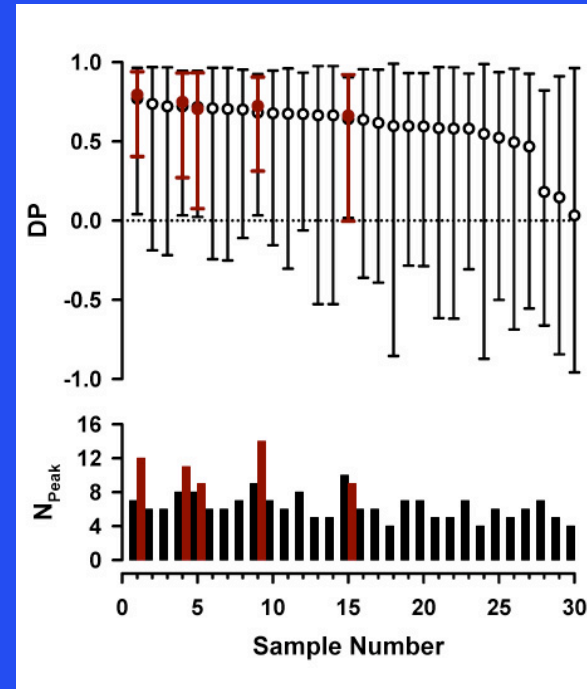
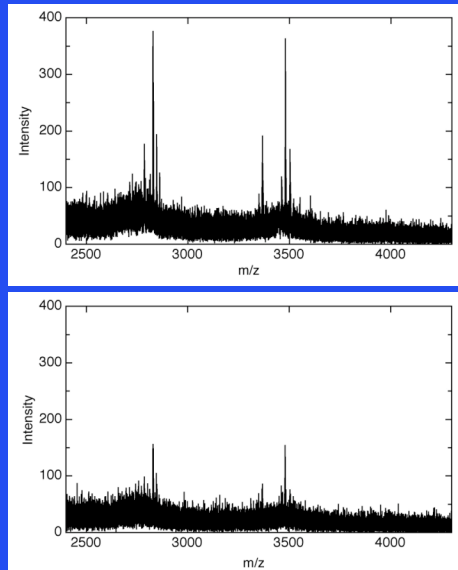
Other 24 replicates very similar to these

FCRC
February 24, 2009

Nature of the Problem - Example 1

Replicate Spectra from Bovine Testicular Tubulin

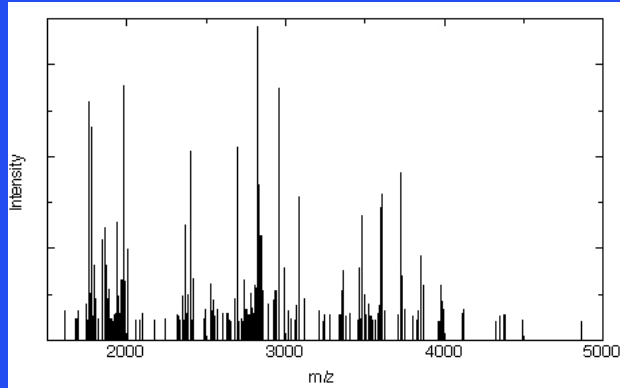
All intensities normalized



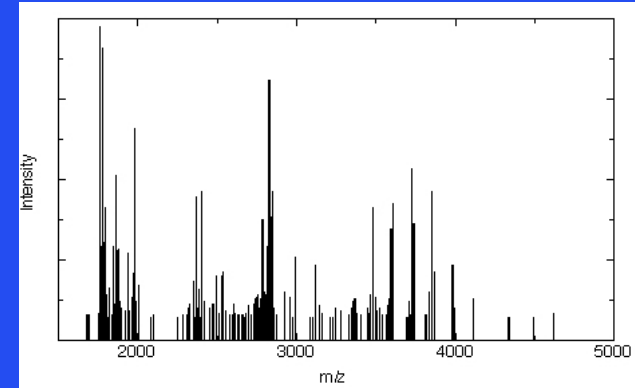
- Only 5 replicates correlated with the consensus spectrum
- Generating Consensus from only the 5 correlating spectra allowed for increased # peaks in the consensus, and much tighter confidence intervals between replicates and consensus

Nature of the Problem - Example 2

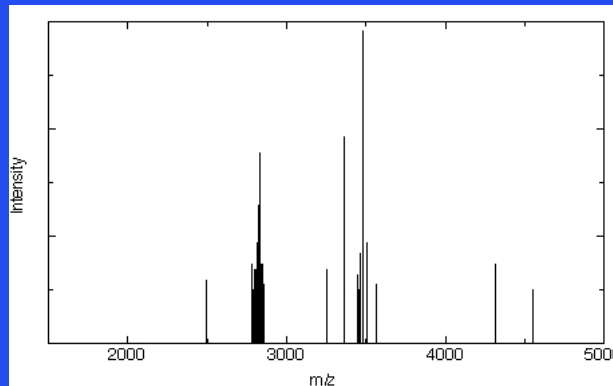
Replicate Spectra of Tubulins from Different Sources -
Similar or Different?
All intensities normalized



Rat Brain



Bovine Brain



Bovine Testicle

Nature of the Problem - Example 2

Replicate Spectra of Tubulins from Different Sources -
Similar or Different?

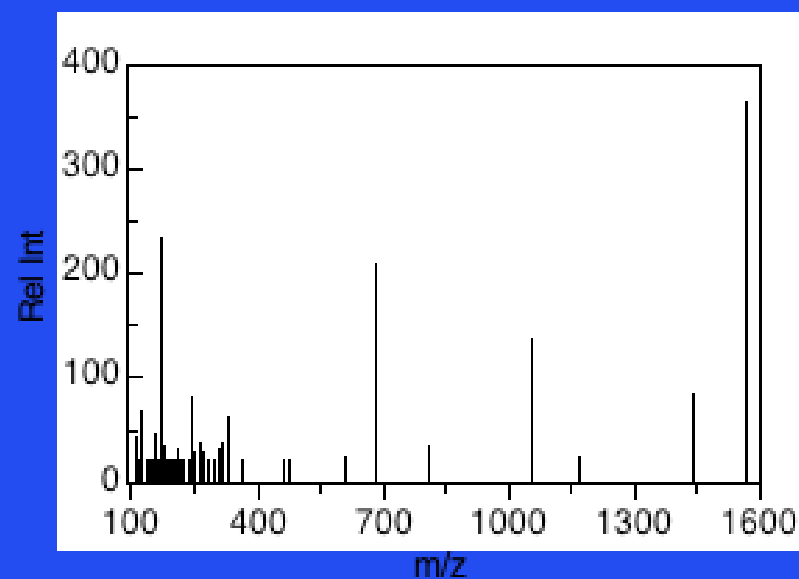
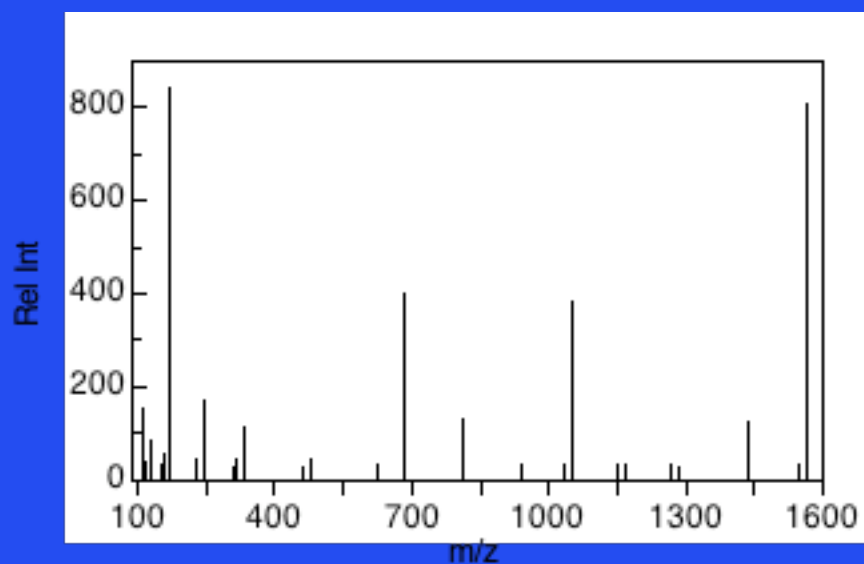
	BBT	BTT
RBT	.645 (.460-.776)	.230 (-.368-.693)
BTT	.187 (-.465-.708)	

BBT - Bovine Brain
BTT - Bovine Testicle
RBT - Rat Brain

Precursor Fragmentation for *de Novo* Sequencing

Replicate Fragmentation Spectra

m/z 1570 - GluFib Peptide



De Novo Sequencing of m/z 1570 Peptide

Generating the Consensus Spectrum

Spectrum	DP vs Consensus	CI vs Consensus
GluFib replicate 1	0.43	-0.11 - 0.77
GluFib replicate 2	0.88	0.69 - 0.96
GluFib replicate 3	0.79	0.42 - 0.94
GluFib replcate 4	0.88	0.70 - 0.96
GluFib replicate 5	0.85	0.57 - 0.95
GluFib replicate 6	0.89	0.72 - 0.96

Evaluating Mass Spectral Similarity and Reproducibility: Does this work? De Novo Sequencing

Spectrum	Sequence: EGVNDNEEGFFSAR	%TIC
GluFib Consensus	E(GNV)DNEEGFFSAR	94.0
GluFib replicate 1	[(DX)][(EV)] <u>GND</u> NEEG[(MY)][(FF)]SAR	81.1
GluFib replicate 2	---	0
GluFib replicate 3	<u>QT</u> SF(MY)E(FG)FAGW	20.8
GluFib replicate 4	E(GNV)DNEEGFFSAR	87.8
GluFib replicate 5	---	0
GluFib replicate 6	<u>EN</u> (GV)DNEEGFFSAR	86.3

Four BSA Peptides Consensus Sequencing de Novo

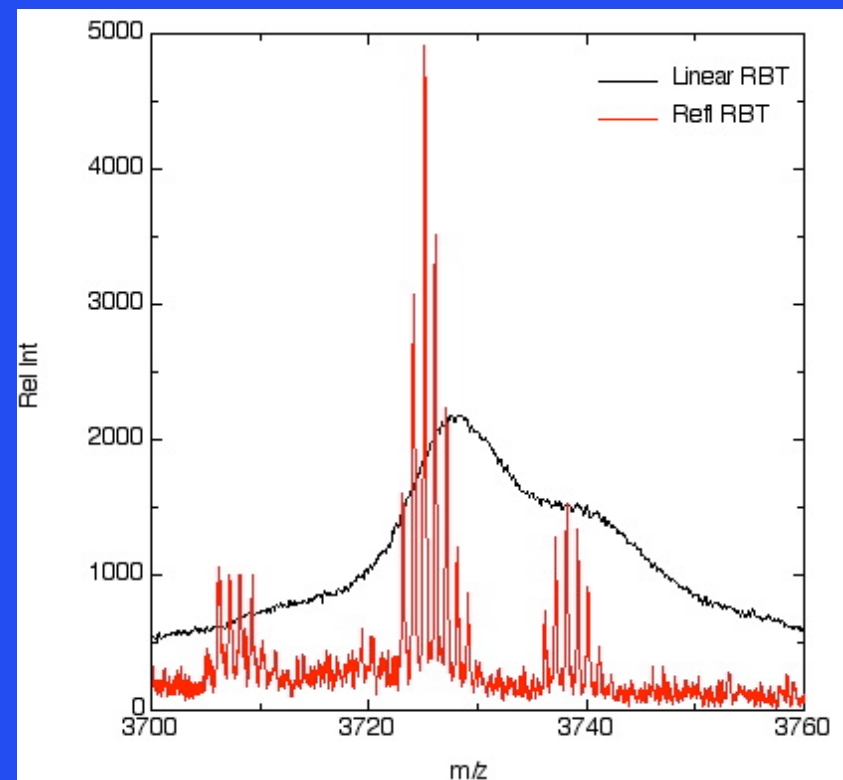
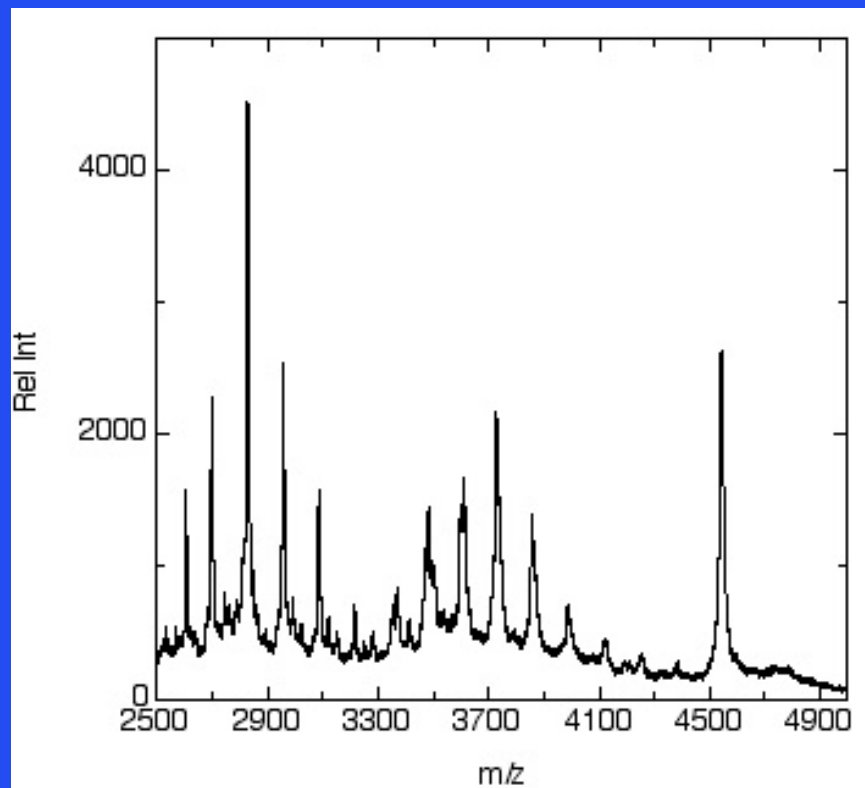
Peptide	Well 1	Well 2	Well 3	18 Reps
YLYEIAR (m/z 927)	YXYEXAR 62% (5/6 Reps)	YXYEXAR 57% (6/6 Reps)	YXYEXAR 61% (6/6 Reps)	YXYEXAR 64%
LGEYGFQNALIVR (m/z 1479)	XGEYGFKNAXXVR 85% (2/6 Reps)	XGEYGFKNAXXVR 82% (4/6 Reps)	(GX)EYGFKNAXXVR 82% (1/6 Reps)	XGEYGFKNAXXVR 84%
DAFLGSFLYEYSR (m/z 1567)	D(AF)XGSFXYEYSR 76% (4/6 Reps)	DAFXGSFYE(HX)R 77% (1/6 reps correct)	DAFXGSFXYEYSR 82% (3/6 Reps)	DAFXGSFXYEYSR 80%
KVPQVSTPTLVEVSR (m/z 1639)	KVPKVST(PT)XVEVSR 77% (3/6 Reps)	KVPKVST(PT)XVEVSR 77% (1/6 Reps)	No hits (1/6 Reps)	KVPKVST(PT)XVEVSR 82%

Can We Extend This Concept To Be More Generally Useful?

QT Clustering

- A “greedy” algorithm from genomics designed to form clusters of genes with each cluster having a minimum level of quality
- Algorithm looks through a list of genes and finds those with the greatest similarity to some initial choice, and keeps hunting until no further matches within the quality threshold can be found.
- Process continues for all genes to form a set of candidate clusters.
- The best cluster, with at least the minimum number of pre-selected components is chosen and its components removed from the list and the process begins again.
- The process continues until all possible clusters are formed.

QT Clustering Applied to Linear MALDI Spectra



Results of QT Clustering

2502.816	3,3,3,4,3,3,3
2513.131	7,5,4,7,4,4,6,5,6,6,4,6,4,7,6,4,6,7,7,3,4
2521.893	3,3,3,3
2529.892	19,21,19,19,19,9,15,19,20,17,18,16,15,17,17,13,15,15,15,16,16,13
2540.277	10,10,10,10,10,13,9,10,6,10,10,9,9,11,8,10,9,9,7,8,8
2569.046	18,18,13,16,18,19,20,18,15,12,19,15,17,15,16,14,14,13,17,14,13,8
2585.349	5,5,6
2586.403	4,5,4,4,4,6,6,6,5,3,4,6,3,8
2604.466	150,70,137,120,112,97,97,90,98,95,141,91,108,103,137,131,83,110,106,124,92,117
2621.487	7,8,8,10,11,10,6,7,7,9,4,9,9,6,7,6,6,6,6
2622.037	7,6,6
2634.802	6,10,6,8,7,8,12,7,7,9,6,6,8,13,14
2635.880	16,14,15,13,14,13,14
2641.270	5,4,4,4,4,3,4,6,5
2658.596	4,4,3,3
2659.226	4,4,4,4,3,4,3,4,4,4

L10-Gel 1_1	0.914	94	0.874	0.942
L10-Gel 1_10	0.924	75	0.882	0.951
L10-Gel 1_11	0.914	94	0.873	0.942
L10-Gel 1_2	0.958	100	0.938	0.971
L10-Gel 1_3	0.952	96	0.929	0.968
L10-Gel 1_4	0.955	95	0.933	0.970
L10-Gel 1_5	0.957	99	0.936	0.971
L10-Gel 1_6	0.959	98	0.940	0.972
L10-Gel 1_7	0.957	98	0.936	0.971
L10-Gel 1_8	0.948	91	0.922	0.965
L10-Gel 1_9	0.942	86	0.913	0.962
L11-Gel 1_1	0.915	94	0.874	0.943
L11-Gel 1_10	0.957	95	0.936	0.971
L11-Gel 1_11	0.894	90	0.843	0.929
L11-Gel 1_2	0.923	98	0.887	0.948
L11-Gel 1_3	0.911	90	0.868	0.941
L11-Gel 1_4	0.939	84	0.908	0.960
L11-Gel 1_5	0.956	101	0.935	0.970
L11-Gel 1_6	0.955	96	0.933	0.970
L11-Gel 1_7	0.961	102	0.942	0.973
L11-Gel 1_8	0.954	96	0.932	0.969
L11-Gel 1_9	0.950	92	0.926	0.967

ECRC
February 24, 2009

Allowing Us to See High Levels of Glutamylation

3537.63	0.73	0.29	RBT a4 cterm + 7E
3538.260		0.29	
3551.63	0.66	0.19	RBT b2 1 MC glob 268 - 299
3552.356		0.17	
3568.633		0.28	
3598.231		1.95	
3609.07	-0.43	2.93	b4a cterm + 1E
3609.95	0.46	2.34	b4a cterm + 1E
3625.39	0.01	0.33	RBT b5 cterm + 2E
3651.841		0.14	
3652.665		0.12	
3666.08	0.14	0.14	RBT a4 cterm + 8E
3667.449		0.12	
3698.781		0.11	
3727.06	-0.21	4.21	RBT b5 glob 331 - 363, b4a glob 331 - 363, RBT b2 glob 331 - 363
3730.27	0.13	6.58	RBT K-a1 cterm - Y + 8E
3738.69	0.15	1.52	b4a cterm + 2E
3791.753		0.14	
3793.446		0.12	
3856.311		2.30	
3859.14	-0.05	3.85	RBT K-a1 cterm - Y + 9E
3868.44	0.86	0.72	b4a cterm + 3E
3968.376		0.12	
3985.746		0.68	
3987.95	-0.28	1.32	RBT K-a1 cterm - Y + 10E
3996.78	0.15	0.21	b4a cterm + 4E
4116.243		0.24	
4117.24	-0.04	0.19	RBT K-a1 cterm - Y + 11E
4119.236		0.75	
4119.792		0.47	
4123.702		0.30	

Summary

- A Consensus Spectrum from complex spectra
 - Allows one to account for variance
 - Permits comparison of spectra from different sources
- Use of the Dot Product
 - Allows use of Pearson's Correlation Coeff
(For Normalized, Mean Centered Spectra)
- The combination can be used in MS and MS/MS

Future Directions

If we look at the “quality”
of hits in a DB search of
LC-MS/MS spectra (.dta files) -

Will the quality improve by
using consensus spectra
vs
the each of the replicates?

Coworkers & Funding

Matthew Olson
Dan Sackett
Paul Blank
Jonathan Epstein

Nancy Vieira
Peter Harrington

NICHD IRP